WEEK 5

Working with Real World Datasets

Rav Ahuja

Working with CSV files

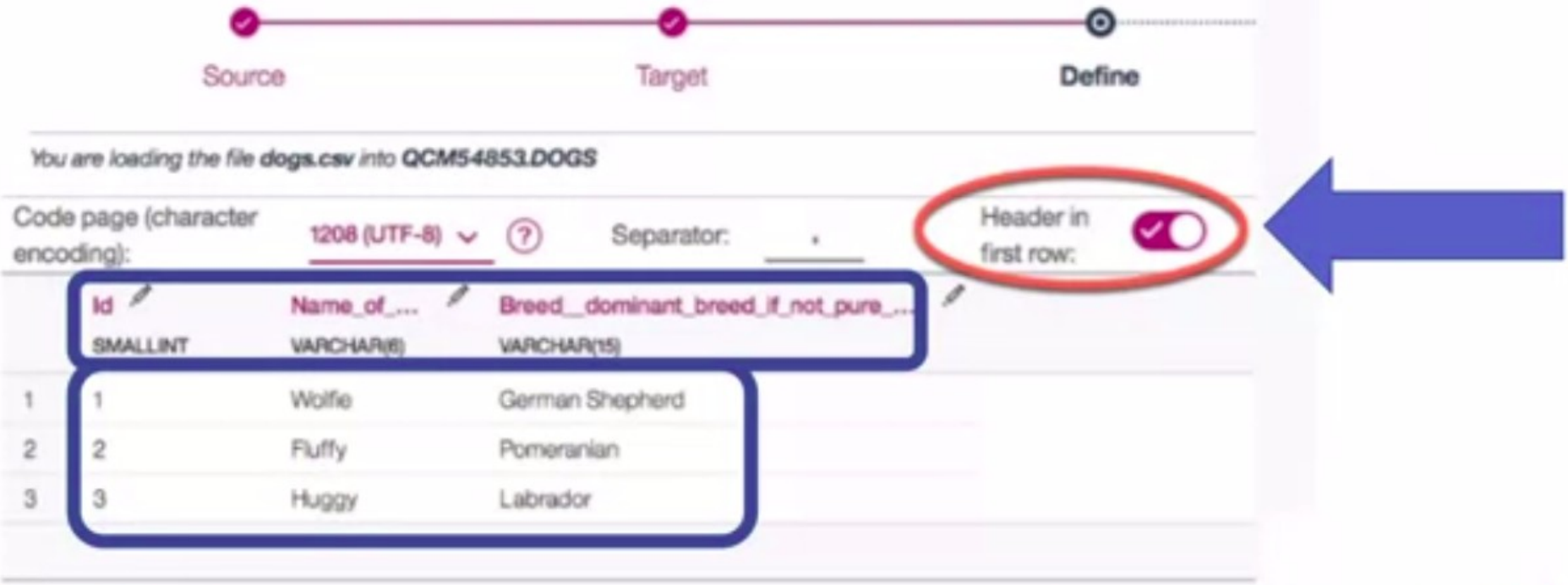
- Many real data sets are .CSV files
- .CSV: COMMA SEPARATED VALUES

- Example: DOGS.CSV

```
Id,Name of Dog,Breed (dominant breed if not pure breed)
1,Wolfie,German Shepherd
2,Fluffy,Pomeranian
3,Huggy,Labrador
```

Column names in first row

When header row in CSV file contains column names:



Querying column names with mixed (upper and lower) case

Retrieve Id column from DOGS table. Try:

```
select id from DOGS
```

If you run this query, you will get this error:

```
Error: "ID" is not valid in the context where it is
used.. SQLCODE=-206, SQLSTATE=42703, DRIVER=4.22.36
```

Use double quotes to specify mixed-case column names:

```
select "Id" from DOGS
```

Querying column names with spaces and special characters

By default, spaces are mapped to underscores:

	A
1	Name of Dog
2	

→ `Name_of_Dog`

Other characters may also get mapped to underscores:

```
select "Id", "Name_of_Dog",
"Breed__dominant_breed_if_not_pure_breed_"
from dogs
```

Breed (dominant breed if not pure breed)

Querying column names with spaces and special characters

By default, spaces are mapped to underscores:

	A
1	Name of Dog
2	

→ `Name_of_Dog`

Other characters may also get mapped to underscores:

```
select "Id", "Name_of_Dog",
"Breed__dominant_breed_if_not_pure_breed_"
from dogs
```

Breed (dominant breed if not pure breed)

Using quotes in Jupyter notebooks

First assign queries to variables:

```
selectQuery = 'select "Id" from dogs'
```

Use a backslash \ as the escape character in cases where the query contains single quotes:

```
selectQuery = 'select * from dogs
      where "Name_of_Dog"=\'Huggy\' '
```

Splitting queries to multiple lines in Jupyter

Use backslash "\" to split the query into multiple lines:

```
%sql select "Id", "Name_of_Dog", \
     from dogs \
     where "Name_of_Dog"='Huggy'
```

Or use %%sql in the first row of the cell in the notebook:

```
%%sql
select "Id", "Name_of_Dog",
     from dogs
     where "Name_of_Dog"='Huggy'
```

Restricting the # of rows retrieved

To get a sample or look at a small set of rows, limit the result set by using the LIMIT clause:


```
select * from census_data LIMIT 3
```

Getting Table and Column Details

Rav Ahuja

Getting a list of tables in the database

Getting a list of tables in the database

DB2
SYSCAT.TABLES

SQL Server
INFORMATION_SCHEMA.TABLES

Oracle
ALL_TABLES or USER_TABLES

Getting a list of tables in the database

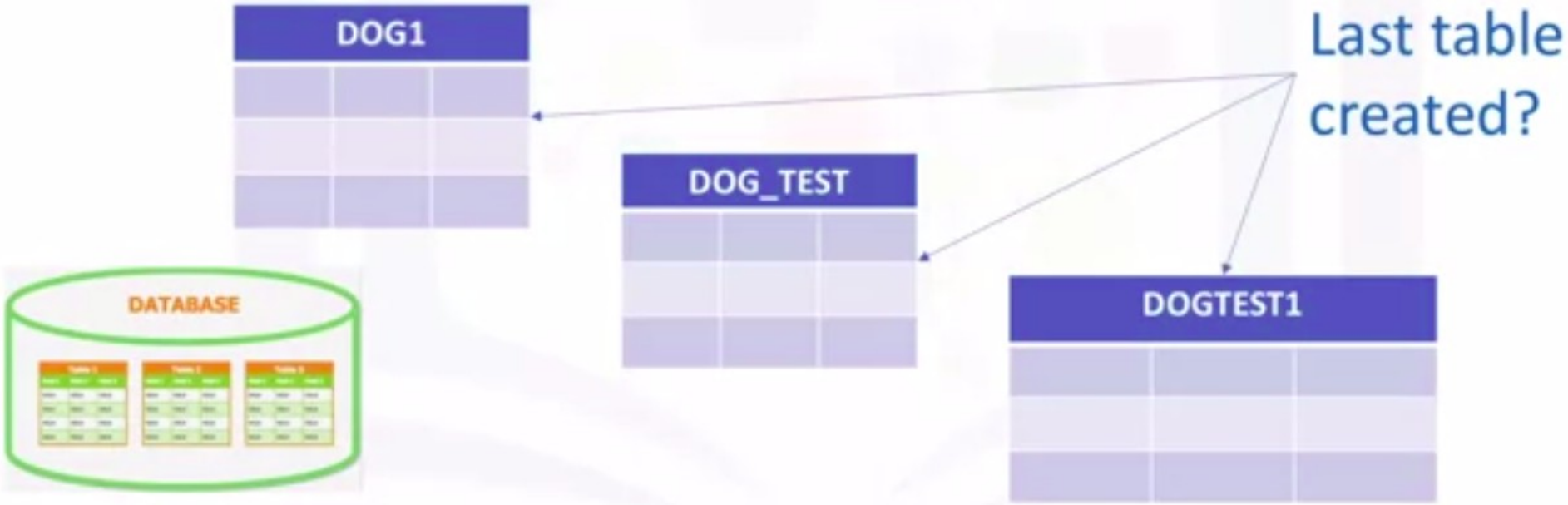
Query system catalog to get a list of tables & their properties:

```
select * from syscat.tables
```


```
select TABSCHEMA, TABNAME, CREATE_TIME
    from syscat.tables
    where tabschema= 'ABC12345'
```

Getting Table Properties

```
select * from syscat.tables
```



Getting Table Properties

```
select TABSCHEMA, TABNAME, CREATE_TIME
     from syscat.tables
     where tabschema='LCT12330'
```

```
[15]: %sql select TABSCHEMA, TABNAME, CREATE_TIME from SYSCAT.TABLES where TABSCHEMA='LCT12330'
 * ibm_db_sa://lct12330:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

[15]:

tabschema	tabname	create_time
LCT12330	PETRESCUE	2020-05-05 22:02:25.169368
LCT12330	BOOK	2020-02-20 22:37:50.351829
LCT12330	MEDALS	2020-02-22 02:50:54.841324
LCT12330	EMPLOYEES	2020-02-18 20:05:38.328981
LCT12330	JOB_HISTORY	2020-02-18 20:05:38.521203
LCT12330	JOBS	2020-02-18 20:05:38.700230
LCT12330	LOCATIONS	2020-02-18 20:05:39.050699
LCT12330	SCHOOLS	2020-04-16 22:06:58.174131
LCT12330	DEPARTMENTS	2020-04-15 20:01:16.429000
LCT12330	MCDONALDSNUTRITION	2020-04-23 16:40:35.205237

Getting a list of columns in the database

To obtain the column names query syscat.columns:

```
select * from syscat.columns
       where tabname = 'DOGS'
```

To obtain specific column properties:

```
select distinct(name), coltype, length
       from sysibm.syscolumns
       where tbname = 'DOGS'
```

Column info for a real table

In [12]:
```
%sql select distinct(name), coltype, length \
     from sysibm.syscolumns where tbname = 'CHICAGO_CRIME_DATA'
```

```
 * ibm_db_sa://qcm54853:***@dashdb-txn-sbox-yp-dal09-04.services.dal
Done.
```

Out[12]:

name	coltype	length
Arrest	VARCHAR	5
Beat	SMALLINT	2
Block	VARCHAR	35
Case_Number	VARCHAR	8
Community_Area	DECIMAL	4
Date	VARCHAR	22
Description	VARCHAR	46
District	DECIMAL	4
Domestic	VARCHAR	5
FBI_Code	VARCHAR	3

Back

Final Exam

Graded Quiz • 30 min • 10 total points

Due Sep 10, 11:59 PM IST

00:19:38 remaining

1. The SELECT statement is called a _____, and the output we get from executing the query is called a result set.

1 point

- ○ Table name
- ◉ Query
- ○ Function
- ○ Operator

2. Which of the following SQL statements will delete the customers where the Country is Italy?

1 point

- ○ DELETE FROM CUSTOMERS WHERE COUNTRY IS 'ITALY'
- ◉ DELETE FROM CUSTOMERS WHERE COUNTRY = 'ITALY'
- ○ DELETE COUNTRY 'ITALY' FROM CUSTOMERS
- ○ DELETE 'ITALY' FROM CUSTOMERS

3. What does the primary key of a relational table do?

1 point

3. What does the primary key of a relational table do? **1 point**

- ○ The primary key uniquely identifies each attribute in a table.
- ○ The primary key uniquely identifies each column in a table.
- ◉ The primary key uniquely identifies each row in a table.
- ○ The primary key uniquely identifies each relation in a table.

4. The basic categories of the SQL language based on functionality are _____ and Data Manipulation Language (DML). **1 point**

- ◉ Data Definition Language (DDL)
- ○ Data Entry Language (DEL)
- ○ Data Input Language (DIL)
- ○ Data Update Language (DUL)

5. When querying a table called Representative that contains a list of representatives and the state that they represent, which of the following queries will return the number of representatives from each state? **1 point**

5. When querying a table called Representative that contains a list of representatives and the state that they represent, which of the following queries will return the number of representatives from each state?

1 point

- ○ SELECT State, count(State) FROM Representative
- ◉ SELECT State, count(State) FROM Representative GROUP BY State
- ○ SELECT distinct(State) FROM Representative
- ○ SELECT State, distinct(State) FROM Representative GROUP BY State

6. You want to retrieve a list of cities in a state that have between 10,000 and 20,000 residents. Which clause would you add to the following SQL statement: **SELECT City, Residents FROMState**

1 point

- ○ WHERE Residents 10000 – 20000
- ○ WHERE Residents ARE BETWEEN 10000 AND 20000
- ◉ WHERE Residents BETWEEN 10000 AND 20000
- ○ WHERE Residents IN (10000, 20000)

7. Which of the following queries will retrieve the LOWEST value of PRICE in a table called PRODUCTS?

1 point

- ○ SELECT MAX(PRICE) FROM PRODUCTS
- ○ SELECT LOWEST(PRICE) FROM PRODUCTS
- ◉ SELECT MIN(PRICE) FROM PRODUCTS
- ○ SELECT LEAST(PRICE) FROM PRODUCTS

8. Which of the following queries will retrieve the PRODUCT NAME that has the highest price?

1 point

- ◉ SELECT PRODUCT_NAME FROM PRODUCTS WHERE UNIT_PRICE = (SELECT MAX(UNIT_PRICE) FROM PRODUCTS)
- ○ SELECT MAX(UNIT_PRICE) FROM PRODUCTS
- ○ SELECT PRODUCT_NAME FROM PRODUCTS WHERE UNIT_PRICE IS HIGHEST
- ○ SELECT PRODUCT_NAME FROM PRODUCTS WHERE UNIT_PRICE = MAX

9. A database cursor is a control structure that;

1 point

Back

Final Exam

Graded Quiz • 30 min • 10 total points

Due Sep 10, 11:59 PM IST

00:19:24 remaining

9. A database cursor is a control structure that;

1 point

- ○ Does not allow you to update records within a database
- ○ Does not allow communication with a database
- ◉ Enables traversal over the records in a database
- ○ Does not allow you to create tables

10. Cell magics: start with a double %% sign and apply to the entire cell. (T/F)

1 point

- ○ False
- ◉ True

Coursera Honor Code Learn more

☐ I, **Tushar Raha**, understand that submitting work that isn't my own may result in permanent failure of this course or deactivation of my Coursera account.

Assignment Preparation: Working with real-world data sets and built-in SQL functions

Assignment

Final Exam

Course Wrap-up

Reading: Congratulations & Next Steps
2 min

Congratulations on completing this Databases and SQL for Data Science with Python course! We hope you enjoyed it.

This course is part of:

- IBM Data Analyst Professional Certificate
- Applied Data Science Specialization
- IBM Data Science Professional Certificate
- IBM Data Engineering Foundations Specialization
- IBM Data Engineering Professional Certificate

Those of you following a Data Engineering track must also complete the Honors module in this course. It contains information on more advanced SQL techniques that will be useful to you as a Data Engineer.

As a next step, you can explore other courses in these programs, starting with:

- Data Engineering and Data Analytics tracks: Data Analysis with Python
- Data Engineering track: NoSQL Fundamentals: Introduction to NoSQL Databases

We encourage you to leave your feedback and rate the course.

Good luck!